DOCUMENT RESUME

ED 148 145 FL 009 107

AUTHOR Oller, John W., Jr.
TITLE Formative Assessment of Learner Progress in the Classroom.
PUB DATE Nov 77
NOTE 17p.; Paper presented at the meeting of the American Council on the Teaching of Foreign Languages (San Francisco, California, November 24-27, 1977)

EDRS PRICE MF-$0.83 Plus Postage. HC Not Available from EDRS.
DESCRIPTORS Communicative Competence (Languages); *English (Second Language); Formative Evaluation; Item Analysis; *Language Instruction; Language Proficiency; Language Skills; *Language Tests; Language Usage; Modern Language Curriculum; *Multiple Choice Tests; Pragmatics; *Second Language Learning; Teaching Methods; Testing; *Test Items; Test Reliability; Test Validity
IDENTIFIERS *Defense Language Institute; Distractors (Tests); English Comprehension Level

ABSTRACT

This paper questions the purpose of testing in second language instruction. Comments are based on an examination of tests used by the Defense Language Institute for students of English as a second language. Two kinds of tests are used: the English Comprehension Level (ECL), used primarily as a basis for setting exit requirements, and "Book Quizzes," based on the training materials used in three levels of classes. The reliability and validity of these tests is discussed at length. The issue of whether and how ECL influences the kinds of testing and teaching that take place in the classroom is examined, and specific test items are presented and analyzed. The use of distractors on a test is questioned: it is felt that, for the students who know the answer, distractors serve no purpose, and that the use of distractors in general corresponds to no other aspect of the curriculum. It is also felt that ECL should correlate with normal English usage tasks, and that the focus of tests, students, instructors, and the entire curriculum should be on the relationship of utterances in the target language to realistic contexts of discourse. (CLK)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED148145

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY.

"PERMISSION TO REPRODUCE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY
John W. Oller
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

Formative Assessment of Learner Progress in the Classroom

John W. Oller, Jr.
University of New Mexico

At several recent conferences, I have begun to ask teachers and administrators, what is the purpose of the testing that goes on in your classroom or in this institution?[1] Many of them seem to cast about for an answer as if there were some mysterious recently discovered reason for testing. Most of them offer answers such as, "To find out what the students know, or don't know." Few of them, in fact none of them, seem to have a clear conceptualization of testing in relation to the total curriculum. Administrators talk about placement of students in classes, or for subsequent programs. Some speak of motivating the students, evaluating the instructor, testing the curriculum, informing the teacher and student concerning some phase of the curriculum, and some even speak of predicting success in university study, technical training and the like.

A few days ago, during a series of talks at the Defense Language Institute at Lackland I repeated the question to three different audiences (several hundred people in all). Why do you test? What are you trying to accomplish with the tests? The responses were as varied as in any other context where the question has been posed. Most people avoided the question. They wanted to explain the nature of the DLI English language training. Instead of saying

FL009107

why they were doing it, they seemed to feel what they were doing would provide its own why. They were anxious to explain the fact that DLI uses two kinds of tests--a test called the English Comprehension Level (ECL) as well as what are called "Book Quizzes" based on the training materials used in three levels of classes. The courses are divided into elementary, intermediate and advanced levels. At the advanced level the thousands of trainees per year (predominantly, adult male Arabs, and Iranians, plus small numbers from other countries) are placed into specialized courses where they learn the terminology and the technology of flying, electronics, mechanics, and so forth.

The ECL test, which I was encouraged to pore over in some detail, consists of 120 items in a multiple choice format. The test is divided into two parts; the first is devoted to items presented in a listening mode with written choices and the second contains items in a reading format. In each part some of the items are devoted to specific points of grammar, others to vocabulary items, all supposedly taught somewhere in the elementary or intermediate curriculum. According to Mr. Smilgin (personal communication) some of the instructors often complain that the sense of vocabulary items as presented in the course books differs from the sense required in the test.

There are, according to Sako (1977) about 80 different forms of the ECL. All of them are statistically equated against some reference population. There was some question at Lackland as to whether the reference population is a group that took the DLI training some time ago or whether the reference population is updated to the immediately preceding group of graduates from the DLI training for each subsequent testing. Some of the items are clearly discrete point in orientation, aiming at particular bits and pieces of grammar and vocabulary, but others are integrative aimed at less analyzable skills such as paraphrase recognition, or the ability to form an appropriate continuation

in a dialog.

But what happened to the question about the purpose(s) of testing? What the DLI does by way of testing is not itself a rationale. It turns out that the ECL has two basic purposes. First, and most important from the viewpoint of the DLI administration, the ECL is used as a basis for setting exit requirements. Personnel who are going on to " high hazard" jobs must pass the ECL with a score of 80 (on some standardized scale not explained to me); persons going into "normal technical training" on the other hand are allowed to pass with a score of 70 or better (on the same scale); and what were described as the "hands-on" "see-and-do" personnel must pass with a score of 60 or better. A second purpose for the ECL, closely related to the first, is to screen candidates between the elementary and intermediate levels on a basis that was not discussed in any detail.

Is this the only purpose that can be discerned for the ECL? Screening exclusively? Is this the only effect that it has? So far we have said nothing about the effect that it has on the students. On the instructors? On the curriculum as it is realized in the classroom context? How does the ECL influence the kinds of testing (the so-called "book quizzes") and teaching that are done in the classroom? Furthermore, what is the evidence that the variance produced by the test (that is <u>variance</u> in the algebraic sense of the sum of squares of the deviations from the mean, divided by the number of subjects tested minus one) is brought about by the instruction and not by incidental experience outside the classroom? Put non-technically, how do we know that the DLI English training actually produces the variablility in scores on the ECL? Similarly, how can we be certain that the variablility in scores is actually due to differences in English proficiency? (We will return shortly to the question concerning the variance produced by the test.)

To answer the questions concerning the effects of the test on the teaching, I spent several hours visiting five classes at the elementary level, five at the intermediate level, and three more advanced specialized classes. I also examined the course books used at the elementary, intermediate, and advanced levels. In no less than seven out of the ten classes observed at the elementary and intermediate levels, the teachers used variants of test items as the basis for some phase of the classroom procedure.

For instance, one teacher at the intermediate level used the following item on the chalkboard to "demonstrate" the contrast in use between used to and be used to:

(1) He used to _____ everyday.

A. swimming

B. swims

C. swam

D. swim

The instructor had just finished explaining that if some form of be (Teacher: "What are the forms of be?" Students: "Is, am, was, were.") appears before the form used to, the choice of verb to follow must be the -ing form, otherwise, it must be the simple infinitival form. The students selected every choice except D (swim) and apparently did not understand the corrections offered by the teacher. Finally, when all the other choices were eliminated, someone said D. The students seemed to have trouble comprehending the sentence "He used to swim everyday." They did not seem to be able to follow the grammatical explanation at all.

In another course the instructor used only multiple choice test items based on isolated and disconnected sentences as the basis for the entire lesson,

Here are some examples:

(2) The Alamo _____ in 1718.

A. built by the Spanish

B. built the Spanish

C. was built by the Spanish

D. was built the Spanish

(3) He _____ by his brother.

A. was told the news

B. told the news

C. told by the news

D. was told by the news

(4) The Wright brothers _____ in the 1800's and died in the 1900's.

A. borned

B. was borned

C. have been born

D. were born

Subsequent items on the same quiz (one of the "book quizzes"?) speak of going somewhere and getting back, being on time, a cadet studying more, a woman feeding a child, men reading the newspaper, some film seen last weekend, and a great many other completely unrelated states-of-affairs. The foregoing items and all of the other ones not given here were undoubtedly motivated by a concern for bits and pieces of syntactic knowledge that the courses aim to instill through pattern practice.

Another type of item that appeared in another fascicle (of the same quiz)

used by the same instructor involved sentence completion.

(5) The purpose of the school

A. to train the student.

B. the student to train.

C. is to train the student.

D. is the student to train.

(6) He can draw a circle

A. better I can.

B. better than I can.

C. better than can I.

D. than better can I.

(7) John said he liked

A. than television better movies.

B. television than better movies.

C. television better than movies.

D. movies television better than.

(8) Our dog Saki likes

A. play with soft rubber a. [sic]

B. with play rubber soft ball a. [sic]

C. to play with a rubber soft ball.

D. play a to with soft ball rubber. [sic]

In numerous classes, and in the intermediate textbook, students were admonished not to confuse similar forms. For example, one instructor wrote the items in care of and in case of on the board. He avowed that though these forms were very similar, they had different meanings and should not be confused. In another class careful and careless were displayed in a like manner.

To be sure the point had been driven home, the instructor asked one of the students whether or not he was a CAREFUL driver. He replied, "No I a careless driver." The attempt to drive out the potential confusion succeeded in driving it in instead. Then students were asked to guess a synonym for the word careless. No one would guess a synonym so the instructor finally told the answer, reckless. I could not help wondering if the students might not notice the similarity of reck- and wreck. Oddly a reckless driver is rarely without wrecks. It also occurred to me that it would be natural for the students in this context to think of the possible form reckful on analogy with careless, careful. But alas, if any bright young man should have followed these leads set before him by the instructional procedure, he would have profited nothing but confusion for his effort.

What then is the effect of the multiple-choice, discrete point orientation of the ELC? It pushes the instructional procedure inevitably toward the same kind of practices. It drives the quizzes in that direction too. It causes the instructors to do what they are getting paid to do--to teach their students to answer items on the ELC. Or at least this is what the instructors are driven to attempt. Whether or not instruction based on such procedures has the desired effect is another matter. We will return to this matter below. In the meantime, let us consider another question related to the purpose of tests.

What is the purpose of the distractors in a multiple-choice item? The term distractor betrays the answer. The purpose of the incorrect choices is to distract, or to trap, confuse, and mislead the unwary, to confuse even more the learner who is already slightly confused. For the student on the other hand who knows the answer, the distractors serve no purpose--unless they possibly draw his attention and make a subtle impression that may mislead

him too in the future.

Now, here is another question. Where else in any instructional procedure, or any educational program, is it the intentional and deliberate aim of the procedure or program to entrap, confuse, mislead, and befuddle the learner? The answer is that there is no other aspect of any curriculum (known to me) that deliberately sets out to confuse the learner. Why should we tolerate in the testing procedures, and in the instructional procedures, methods that intentionally set out to put the student's mind in a confused disarray? Are there no other methods available? Of course, there are other methods--we'll get back to some of them shortly, but before that, let's return to the question of variance.

What produces the variance in the ELC? Is it really due to differences in learner abilities in English? Is it actually attributable to intentional instructional procedures? No one knew the answers to these questions, but it was possible to determine that the variance in the ECL is at least quite reliable--above .9. That is, the various forms produced from the computer library at the DLI tend to correlate with each other at about the .9 level. This fact can be interpreted statistically as meaning that about 81% of the total variance in the ECL is common to any two forms of the test which happen to be selected. Further, we can conclude that about 19% of the variance is unreliable and is not common to the various forms of the test. Now the question that must be answered is how much of the reliable variance in the ECL (that is, the 81%) is actually attributable to variance in language ability.

One of the instructors in the audience at one of the meetings on August 18 and 19 said that she did not believe that the ECL is sensitive to genuine differences in skill that result from experiences outside the classroom--e.g.,

dating women in the San Antonio area who speak English,visiting with English speaking people on and off the base, etc. Another instructor argued that the test does not assess the learner's ability to talk and therefore is not sensitive to one of the major instructional efforts of the DLI.

Ultimately, the questions that have been and will continue to be raised are not subject to answer by vote. No amalgamation of opinions, regardless how eclectic and appealing they may seem, can tell how much of the reliable variance in the ECL is actually due to real differences in language proficiency. The question can only be answered by careful research, and then only within certain confidence intervals based on statistical probabilities. However, the faint-hearted should take courage, because the research in language testing is very revealing on the whole, considerably more so than in other areas of educational endeavor. Attainable reliabilities and validities are generally higher than in other areas of social and psychological study. Hence, the answers can be had with a rather high degree of certainty.

A pertinent research question is whether the variance in language use tasks that will subsequently be required of DLI graduates is common to the ECL. That is, to what extent does the ECL correlate with (i.e., produce overlapping variance with) normal English use tasks. What is the correlation with ability to read with comprehension the sorts of material the DLI trains people to read? What is the correlation with listening comprehension tasks that the DLI trains people to perform? What is the correlation with speaking tasks that the DLI trains people to carry on? What is the correlation with writing tasks that the DLI trains people to do? The research problem is difficult but not insurmountable, if taken as a research problem. On the other hand, the problem is probably insoluble if approached on the basis of speculation and opinion--

there are simply too many conflicting expert and non-expert opinions.

There are other research angles. It is possible to examine the correlation between the ECL, the "book quizzes" and other testing techniques that have been widely researched with adult populations--e.g., cloze, dictation, oral interview, essay writing, and the like. Discourse processing tasks of all sorts (especially the sorts required of graduates of the DLI) should be examined in relation to the ECL, the "book quizzes", and the entire instructional program. Chances are very good that the ECL can be improved purely as a measurement device by careful study. Furthermore, there is no doubt that it can be improved from the vantage point of instructional procedure and the effect that it is obviously having there. Does the test require people to do the sorts of things with language that are known to produce language acquisition? If not, the test should be modified immediately on the grounds of construct validity. Does the test correlate with other tests of known validity that purport to assess the same kinds of skill(s) the ECL tries to assess? If not, the test should be revised on the grounds of concurrent validity. No one should urge without study that the test be immediately discarded, but this is a possibility for consideration depending on the outcome of the needed research.

"But," someone says, "the procedure is obviously working. The DLI is getting results. We know that graduates are succeeding in subsequent courses of training elsewhere and are not apparently hindered by a lack of English knowledge. How can you argue with success?" There is a way--simply provide more of it, or provide it at a lesser cost. Is there any reason to believe that an already successful program connot be made more successful? The fact is that at present, there are only hunches as to what is producing

the success. Can anyone deny the desirability of finding out just which (if any) aspects of the training are bringing about the observed changes in language proficiency over the course of time? One of the administrators told me that he had been urging for twelve years that the DLI do away with the elementary and intermediate ESL training and begin immediately with the specialized training related to specific curricular objectives that only incidentally require English skill--e.g., teaching flying, electronics, mechanics, etc. Who knows whether or not this might in fact be a vast improvement over the present set up?

Let's get back to the question with which we began. What is the purpose of testing? It would seem that there is substantial evidence right on base at the DLI that the tests have the following effects: (1) They serve as the nearly exclusive basis for the evaluation of student, instructor, and curriculum. (2) They motivate the learner, teacher, and courses by directing attention to the kinds of performances required by the tests. In a large part, the tests define the instructional procedure. Viewed from the overall perspective, testing and teaching would seem to be two sides of the same instructional coin.

Now that we have determined what the actual effects and therefore the reasonable purposes of the tests are, what should we require of the tests in order to maximize the benefit of their effects and minimize the potential damage they might otherwise do. First, the tests must be reliable. The possibility of dropping a plane on a runway prematurely for lack of English comprehension is too great a risk to run. Besides reliability there is the prerequisite of validity. If the test says it measures English comprehension, or English speaking ability, or whatever, if it is valid it must indeed reliably measure exactly what it is supposed to measure. Third, the test must be practical--whether it is to be used for thousands or for tens. Fourth, it must fit into the instructional program smoothly and effectively.

The kinds of tests that meet the foregoing requirements are ready to hand. They can be had in a multiple choice format for wide scale institutional testing, or they can be had in open-ended formats for instructional use. For reasons given above, the multiple choice format is not recommended for instructional (in-the-classroom) use. Among the techniques known to work well is just about any discourse based task that requires (1) the temporal processing of sequences of units in the target language, and (2) the pragmatic mapping of those sequences of linguistic units (or linguistic contexts) onto extralinguistic contexts. Examples include cloze procedure over discourse texts either written or oral (already used in several natural forms at the DLI, especially in the advanced classes), dictation of discourse texts (and elicited imitation, also both used now in a somewhat unsystematic way), essay writing, and simple talk about known contexts or facts. (For further exemplification and explanation see Oller, in press, and the references cited there.)

A Personal Epilog

It is impossible for me to conclude this paper without urging a strong recommendation for change upon the staff of the DLI. What I have to recommend is not only applicable there, but everywhere that languages are taught and in every aspect of school curricula. I have observed the frustration of teachers trying to create the script, the backdrop, and the characters necessary to illustrate meaningless pattern drills. They are in the position of a writer, producer, cast, and crew trying to dramatize the impossible. I have also seen with my own eyes and sensed in my own viscera the frustrations of the students who are trying to guess the inadequately portrayed contexts. When nonsense is the basis of an instructional procedure as it always is in discrete point oriented language teaching, the world changes from sentence to sentence.

The facts change, the persons change, the context changes. No one but a native speaker can understand suc nonsense, and no one who does not already know the language can learn from such nonsense except arduously and with many errors.

There is another way. Otto Jespersen recommended it in 1904. Oller (1963-65) exemplified it in a Spanish program that carries the novice from zero proficiency to near adult native competence, all in a classroom setting. Finocchiaro (1977) at this same institute recommended a version of the pragmatice approach I want to suggest. Instead of focussing attention on the surface form of utterances without regard for the way that surface form is governed by meaning, every aspect of the curriculum should serve to show the relationship between the surface forms of language and the sense of discourse. Of course, this is logically impossible if there is no context to talk about. Therefore, from the very first day of instruction the focus of attention should be on sensible, realistic, discourse contexts--simple ones at first and more complex ones later on. Never should attention be devoted exclusively to meaningless surface form (except perhaps in pronunciation drills?).

Unfortunately, the focus of attention in the elementary and intermediate curriculum of the DLI at Lackland is almost always on the surface form with a systematic disregard for meaning. No one, no matter how deranged or disturbed would ever be apt to say the things that are found in the typical pattern drills used at both of these levels. At the more advanced levels, in the specialized training courses, the focus changes to meaning. Why not put the focus on meaning in the first place? For the cost conscious person, perhaps an estimate on possible increase in effectiveness would whet the necessary appetites. It is my guess (a researchable guess) that the DLI could cut the time

necessary to get to the present levels of attainment by a third, or perhaps in half, by revising the curriculum along the lines that I am suggesting. Furthermore, the students and teachers alike would suffer much less in the process of acquiring the requisite English skills. If the training program were kept at its present length, the product output would probably be vastly improved--by any reasonable measure of language proficiency (including the ECL).

Summing it up, the focus of the tests, the students, the instructors, and the entire curriculum should be on the relationship of utterances in the target language to realistic contexts of discourse.

References

Finocchiaro, Mary. 1977. Some thoughts on the functional-notional syllabus. Paper presented at the Lackland Air Force Base English Language Center-Defense Language Institute. April 21, 1977.

Jespersen, Otto. 1904. How to teach a foreign language. London: Allen and Unwin.

Oller, J. W. Sr. 1963-65. El español por el mundo. Chicago: Encyclopedia Britannica Films.

Oller. J. W. Jr. In press. Language tests at school: a pragmatic approach.

Sako, Sydney. 1977. The ECL testing system for the Defense English Language Training Program. Paper presented at a meeting on testing jointly sponsored by TESOL and AILA, Miami, Florida.

Notes

[1]This paper was prepared as one of a series of four two-day lecture tours sponsored by the TESOL Organization. I am indebted to Mary Finocchiaro whose printed lecture was made available to me by DLI staff. I should also thank Mr. Paul Allen, Mr. Victor Smilgin, and the other gracious members of the DLI staff who hosted my visit in San Antonio. Finally, a debt of sincere gratitude is due Dr. Alatis and Ms. Carole Leclaire for making the advance arrangements and keeping me on schedule. The contents of this paper, for better or for worse, are my own doing.